WATSS: a Web Annotation Tool for Surveillance Scenarios

Federico Bartoli[1], Lorenzo Seidenari[1], Giuseppe Lisanti[1], Svebor Karaman[1,2] and Alberto Del Bimbo[1]

[1]{firstname.lastname}@unifi.it, University of Florence

[2]svebor.karaman@columbia.edu, Columbia University

ABSTRACT

In this paper, we present a web based annotation tool we developed allowing creating collaboratively a detailed ground truth for datasets related to visual surveillance and behavior understanding. The system persistence is based on a relational database and the user interface is designed using HTML5, Javascript and CSS. Our tool can easily manage datasets with multiple cameras. It allows annotating a person location in the image, its identity, its body and head gaze, as well as a potential occlusion or group membership. We justify each annotation type with regards to current trends of research in the computer vision community. We further detail how our interface can be used to annotate each of these annotations type. We conclude the paper with an usability evaluation of our system.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Collection; H.5.2 [**User Interfaces**]: Graphical user interfaces (GUI)

General Terms

Experimentation

Keywords

Annotation, Surveillance

1. INTRODUCTION

The computer vision and pattern recognition community is always seeking more challenging and realistic datasets to work on. Such datasets have been the main driver of recent major advancements in machine learning and pattern recognition. Challenges, associated with public datasets have also pushed researchers to develop methods to go beyond the state-of-the-art. PASCAL VOC [10] had been and is still advancing the accuracy of object recognition, detection and segmentation. A major break-through in image recognition has been recently made possible thanks to the large ImageNet taxonomy [7] allowing to train a deep convolutional neural network with a sufficient amount of data [12].

Recently, researchers started to address the problem of group behavior understanding. Collective behavior understanding, like standing in groups or queuing up has been addressed in [1, 6]. The problem of person to person interaction has been tackled in [3, 17] showing that modelling social behavior can improve tracking performance. Dataset to study group behavior will often be recorded in mildly crowded environments therefore knowing whether a body is fully visible or partially occluded allows to evaluate how the methods are able to cope with occlusion. Moreover, one of the most important social cue is gaze, usually defined as a coarse gaze by the head pose since it is often not possible to detect the real gaze of a person from far field camera.

Dataset annotation is a time consuming and expensive task to perform. Recently large datasets have been annotated with crowd sourcing. Crowd sourcing usually relies on platforms like Amazon Mechanical Turk (AMT), where "turkers" are paid to perform annotations. To properly exploit AMT web based annotation interfaces [18,20] are needed.

In this paper we present an open-source tool we have developed to annotate the MuseumVisitors dataset [2]. This dataset of person and group behavior understanding, can be used for tracking, detection and coarse gaze estimation. We recorded this dataset at the National Museum of Bargello in Florence, Italy as part of the MNEMOSYNE project [11]. We designed the tool as a web application in order to easily gather annotations from multiple users and to allow concurrent annotations. The tool had to deal with multiple kinds of information thus needing a user interface designed specifically for the task. Annotators can insert groups and people identities, gaze and body occlusion.

2. RELATED TOOLS AND DATASETS

In this section we first review some publicly released annotation tools and then discuss the related datasets limitations that triggered the development of the WATSS tool.

2.1 Annotation tools

The LabelMe annotation tool [18] is focused on annotating scenes providing web based tools and mobile applications to annotate, using polygons, the outline of objects. Tools to annotate surveillance videos have been recently proposed such as VIPER [13] and VATIC [20]. These tools usually support annotations like bounding boxes, polygons and ellipses, as they are mostly developed for object detection. VATIC allows to specify a finite set of attribute per every object such as "walking" for "person" objects. The main drawback of a tool like VIPER is that is meant to be used locally instead of

MM'15 October 26-30, 2015, Brisbane, Australia.

online, therefore the gathering of annotations from multiple sources can become difficult and there is no way of connecting the tool with crowdsourcing platforms. VATIC is a more modern online tool that can be used for crowdsourcing at scale, although their data model is extremely focused on detection and structured detection of objects [22]. The possibility to add attributes gives some flexibility to the data model but is not enough to manage the diversity of data needed for behavior understanding.

2.2 Group and occlusion detection datasets

Person detection is widely studied in literature and many datasets have been publicly released, each one with different characteristics. However, there is a lack of datasets with group annotation, that can be used for example in group detection, tracking and behavior analysis. Moreover, very few datasets have gaze annotation. In this section we briefly review some currently available datasets that contain groups or occlusion annotations.

Group detection.

The CAVIAR dataset [5] was released in 2003 for behavior analysis purposes. It consists of two sets of experiments, each one composed by a set of video clips taken from different cameras. These sequences were recorded acting out different scenarios of interest for different behaviors. It comes with groups annotations and it can be exploited for group detection, tracking or behavior analysis.

The Friends Meet (FM) dataset was recently proposed in [3] specifically for group detection and tracking. It contains groups of people that evolve, appear and disappear spontaneously, and experience split and merge events. It is composed of 53 sequences, for a total of 16286 frames. The sequences are partitioned in a synthetic set without any complex object representation and dynamics, and a real dataset.

Occlusion detection.

Recently a lot of techniques have been focusing on person detection with occlusions handling [14,16,21]. However, due to the lack of datasets with occlusion annotations it is difficult to produce a quantitative measure of this phenomenon and compare with other methods. The Daimler Pedestrian Detection Benchmark dataset [9] is a set of images captured from a vehicle-mounted calibrated stereo camera rig that is moving in an urban environment. It contains bounding boxes annotations for pedestrians and non-pedestrians in the scene. No additional annotation are provided about visible (or occluded) part of each pedestrian. However, the test set is split between non-occluded and partially-occluded. The Caltech dataset [8] is composed of 250000 frames extracted from 10 hours of videos acquired from a vehicle driving through regular traffic in an urban environment.

3. WATSS ANNOTATION TOOL

Most of current datasets are targeted for a single task, such as: person detection with occlusion, group detection and/or behavior analysis. Moreover, to the best of our knowledge no open source annotation tools are available to easily produce all the annotations needed to build a dataset covering jointly all these tasks.

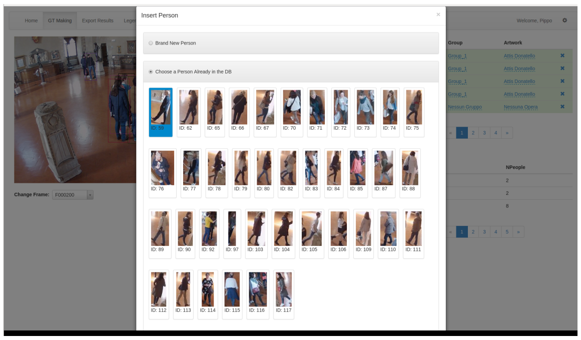
We hence developed a web-based annotation tool to annotate our MuseumVisitors Dataset [2] and we made the source code publicly available. This dataset is a great example of what is needed in a modern visual surveillance dataset. In our case we want as much information as possible so we developed functionalities to annotate position, person identity, gaze, occlusion persons and group membership.

3.1 Annotation protocol

We propose the following annotation protocol. First of all people bounding boxes must be defined, a bounding box can be positioned and rescaled to better fit a person. If a person is partially occluded, a secondary bounding box annotation corresponding only to the visible part of the person can be defined as shown in Fig. 3(a).

Annotators can provide identities for pedestrians associating a single identifier on all frames of all cameras. Identities are easily assigned thanks to our *Add person* interface showing avatars of already enrolled identities as show in Fig. 2.

In presence of groups, annotators can also associate a group identifier that is common to all frames of all cameras. Finally, it is possible to specify body orientation and gaze with a quantization of 5 degrees as shown in Fig. 3(b).

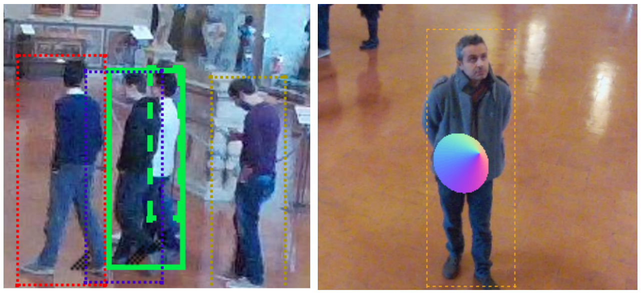


Change Frame: F000200

Group	Artwork	
Group_1	Attis Donatello	✖
Group_1	Attis Donatello	✖
Group_1	Attis Donatello	✖
Group_1	Attis Donatello	✖
Nessun Gruppo	Nessuna Opera	✖

NPeople
2
2
8

Figure 2: Add person view. Annotators can add a new identity or select from a one previously inserted.

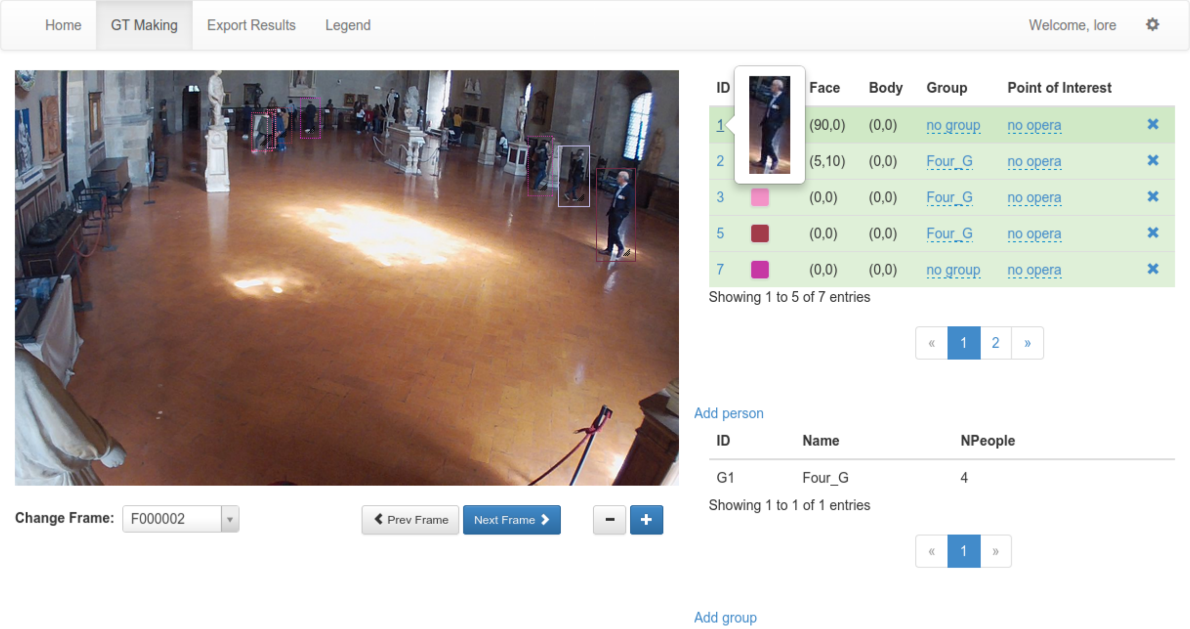


(a) (b)

Figure 3: *(a)* The solid green rectangle represent the bounding box selected for the annotation while the green dashed rectangle represent the visible (not occluded) area annotated by the user; *(b)* The cone visualizes the annotation of the gaze provided by the user.

3.2 The web based annotation tool

We designed a user friendly web interface to ease the tedious task of a detailed surveillance videos annotation. Im-



ID		Face	Body	Group	Point of Interest	
1		(90,0)	(0,0)	no group	no opera	✖
2		(5,10)	(0,0)	Four_G	no opera	✖
3	🟪	(0,0)	(0,0)	Four_G	no opera	✖
5	🟥	(0,0)	(0,0)	Four_G	no opera	✖
7	🟪	(0,0)	(0,0)	no group	no opera	✖

Showing 1 to 5 of 7 entries

« 1 2 »

ID	Name	NPeople
G1	Four_G	4

Showing 1 to 1 of 1 entries

« 1 »

Figure 1: Web interface. Showing several persons bounding boxes and the pop-up avatar for the first one.

plementing the tool as a web platform allows concurrent annotation. In fact, multiple annotators can be easily tasked with a different range of frames to annotate. Moreover the interface implement a function to point an annotator to the next un-annotated frame. In Fig. 1 we show the interface.

On the top of the interface we have a menu bar with different options: *GTmaking*, *Export results* and *Legend*. If a user selects *GT making* the annotation tool asks for username and allows to chose the camera and frame to annotate, if none is specified the annotation process will start from the latest frame annotated by the user.

On the *left-top* part of the interface, we show the chosen frame along with some already annotated bounding boxes. By selecting one of the bounding boxes the dashed rectangle become solid and the user is able to move and resize the bounding box. Once a bounding box is selected the user can also specify different information about that annotation, such as: the visible area (occlusion), the direction of the body and the gaze. A new bounding boxes can be added by clicking "Add person".

On the *left-bottom* part of the interface, we put some video related buttons that allows to navigate through the frames and zoom-in or out on the image (annotators can zoom also by scrolling with the mouse or touchpad).

In the *right-top* part of the interface we put one table summarizing the information about each individual, like the person identifier (ID), the color of the bounding box, the gaze direction (Face), the body direction (Body), the group of which the selected user is part of (Group) and if it is standing by a particular object in the scene or not (Object).

In the *right-bottom* part of the interface we put, instead, a table summarizing the groups information, like the identifier of the group (ID), the name of the group (Name) and the number of persons that are part of the group (NPeople). A new group can be added by clicking "Add group" below the table.

The tool now supports CSV exporting, clicking on export data triggers the generation of an archive containing the CSV files with the annotated data.

In order to make this tool intuitive and ease the annotation process we defined a series of keyboard shortcuts to speed-up the process. These shortcuts are summarized in the *Legend* section of the annotation tool. Moreover, once a frame is annotated, the successive frame will have the same bounding boxes as a starting point for the new annotations, in order to overcome the necessity of re-defining from scratch every person annotation at every frame.

3.3 Usability evaluation

To evaluate the usability of the proposed annotation tool we used the System Usability Scale (SUS) [4], which is a Likert scale. The form to create a Likert [19] scale is built by presenting a set of questions and asking the respondent to choose a degree of agreement in a fixed point scale, from strongly disagree to strongly agree (in our case 1 to 5). It is not just a forced choice questionnaire. Questions are selected in order to present extreme cases and alternating positive and negative statements. The alternation of positive and negative statements is a way of making sure that the respondent reads carefully. The selection of extreme scenarios is instead a way of removing bias. The SUS questionnaire was build selecting among a pool of 50 questions, those leading to the most extreme responses.

We report our usability study result in Table 1. As suggested by Nielsen [15] five system users are enough to find the 85% of usability issues of interfaces. Regarding the SUS score our system obtained an average score of 70. We noted that all users found to be confident using the system (item 10), and the system easy to learn and use (items 2,3). We also found that many user gave a neutral response to item 6; this is probably caused by the diversity of annotations requested, but it is also room for improvement.

4. CONCLUSION

We presented a web annotation system designed for annotating multi-camera video sequences typical of surveillance scenarios.

	Str. Dis.	Dis.	Neutr.	Agr.	Str. Agr.
1. I think that I would like to use this system to perform an annotation task	0	0	2	2	1
2. I imagine that most people would learn to use this system very quickly	1	**3**	0	1	0
3. I found the system very cumbersome to use	2	2	0	1	0
4. I thought the system was easy to use	0	1	0	**4**	0
5. I think that I would need the help of a technical person to use this system	2	2	1	0	0
6. I found the various functions in this system were well integrated	0	0	**4**	1	0
7. I thought there was too much inconsistency in this system	2	2	0	1	0
8. I found the system unnecessarily complex	2	2	0	1	0
9. I needed to learn a lot of things before I could get going with this system	2	2	0	0	1
10. I felt very confident using the system	1	0	1	2	1

Table 1: Result of our SUS usability study. We report frequencies of each answers. Most frequent items are reported in bold.

We tested WATSS annotating our publicly released MuseumVisitors dataset comprised of 96972 detections, and gazes, 101 persons' identities over 9477 frames from four cameras. This is the work of 5 people performed through our interface for 20 days: roughly 3 man/months. We evaluated the system usability using the well known SUS scale finding that the system is considered easy to learn and use and annotators felt productive and confident in using it.

The tool is available on bitbucket at https://bitbucket.org/fbert/watss[1] under GPLv3 License. We provide installation scripts to feed frames into the system that can be tested at http://150.217.35.152/watss. We release our MuseumVisitors dataset together with the tool so that annotations can be visualized on a real world scenario.

With respect to a tool like VATIC we have a specific interface to annotate occlusions and user gaze. Moreover we are able to easily annotate user identity by showing the annotator previous persons frames. Our system provides suggestions for bounding boxes and gazes for subsequent frames so that annotators have to perform a simpler tuning task instead of redefining all scene entities from scratch. Considering the complexity of the scenarios usually involved we are not able, at the moment, to allow the interpolation of coarsely annotated sequences via tracking as in [20]. We plan in the future to add more sensible proposals for unannotated frames both for gaze and detections in order to reduce the complexity of the annotation process.

Acknowledgments. Lorenzo Seidenari is supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST.

5. REFERENCES

[1] M. Amer, P. Lei, and S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proc of ECCV*, 2014.

[2] F. Bartoli, G. Lisanti, S. Seidenari, Lorenzo Karaman, and A. Del Bimbo. Museumvisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding. In *Proc. of CVPR Int.'l Workshop on Group And Crowd Behavior Analysis And Understanding*, Boston, USA, 2015.

[3] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *Proc. of CVPR*, 2012.

[4] J. Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[5] CAVIAR. Test case scenarios. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proc. of ECCV*, 2012.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of CVPR*, 2009.

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, April 2012.

[9] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Proc. of CVPR*, 2010.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[11] S. Karaman, A. D. Bagdanov, L. Landucci, G. D'Amico, A. Ferracani, D. Pezzatini, and A. Del Bimbo. Personalized multimedia content delivery on an interactive table by passive observation of museum visitors. *Multimedia Tools and Applications*, pages 1–25, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*. 2012.

[13] V. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, D. Doermann, and T. Drayer. Performance evaluation of object detection algorithms. international conference on pattern recognition. In *In Proc. of ICPR*, 2002.

[14] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool. Handling occlusions with franken-classifiers. In *Proc. of ICCV*, 2013.

[15] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, pages 249–256, New York, NY, USA, 1990. ACM.

[16] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *Proc. of CVPR*, 2013.

[17] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. of ICCV*, 2009.

[18] B. C. Russell, A. A. Torralba, and F. W. T. Murphy, K. P. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, May 2008.

[19] W. M. Trochim et al. Likert scaling. *Research methods knowledge base*, 2, 2006.

[20] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21.

[21] P. Wohlhart, M. Donoser, P. M. Roth, and H. Bischof. Detecting partially occluded objects with an implicit shape model random field. In *Proc. of ACCV*, 2012.

[22] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.

[1] Direct download: https://goo.gl/cgihhr